

The Latin American Giant Observatory: a successful collaboration in Latin America based on Cosmic Rays and computer science domains

H. Asorey^{*1}, R. Mayo-García^{†2}, L. A. Núñez^{‡3,4}, M. Rodríguez-Pascual^{§2}, A. J. Rubio Montero^{¶2}, M. Suarez-Durán^{||3}, L. A. Torres-Niño^{**4}, and for the LAGO Collaboration^{††5}

¹Laboratorio Detección de Partículas y Radiación, Instituto Balseiro & Centro Atómico Bariloche, San Carlos de Bariloche, Argentina

²División de Tecnologías de la Información y las Comunicaciones, Centro de Investigaciones Energéticas Medioambientales y Tecnológicas, Madrid, Spain

³Escuela de Física, Universidad Industrial de Santander, Bucaramanga, Colombia

⁴Centro de Supercomputación y Cálculo Científico, Universidad Industrial de Santander, Bucaramanga, Colombia

⁵lagoproject.org, see the full list of members and institutions at lagoproject.org/collab.html.

May 31, 2016

Abstract

In this work the strategy of the Latin American Giant Observatory (LAGO) to build a Latin American collaboration is presented. Installing Cosmic Rays detectors settled all around the Continent, from Mexico to the Antarctica, this collaboration is forming a community that embraces both high energy physicist and computer scientists. This is so because the data that are measured must be analytical processed and due to the fact that *a priori* and *a posteriori* simulations representing the effects of the radiation must be performed. To perform the calculi, customized codes have been implemented by the collaboration. With regard to the huge amount of data emerging from this network of sensors and from the computational simulations performed in a diversity of computing architectures and e-infrastructures, an effort is being carried out to catalog and preserve a vast amount of data produced by the water-Cherenkov Detector network and the complete LAGO simulation workflow that characterize each site. Metadata, Permanent Identifiers and the facilities from the LAGO Data Repository are described in this work jointly with the simulation codes used. These initiatives allow researchers to produce and find data and to directly use them in a code running by means of a Science Gateway that provides access to different clusters, Grid and Cloud infrastructures worldwide.

Keywords: Cosmic rays; big data; Corsika; HPC; LAGO

1 Introduction

The LAGO (Latin American Giant Observatory) project is an extended Astroparticle Observatory at global scale. It is mainly oriented to basic research on three branches of Astroparticle physics: the Extreme Universe, Space Weather phenomena, and Atmospheric Radiation at ground level.

^{*}asoreyh@cab.cnea.gov.ar

[†]rafael.mayo@ciemat.es

[‡]lnunez@uis.edu.co

[§]manuel.rodriguez@ciemat.es

[¶]antonio.rubio@ciemat.es

^{||}mauricio.suarez@correo.uis.edu.co

^{**}alejandro.torres@correo.uis.edu.co

^{††}лаго@lagoproject.org

The LAGO detection network consists in single or small arrays of particle detectors at ground level, spanning over different sites located at significantly different latitudes (currently from Mexico up to the Antarctic region) and different altitudes (from sea level up to more than 5,000 meters over sea level), covering a huge range of geomagnetic rigidity cut-offs and atmospheric absorption/reaction levels.

The LAGO Project is operated by the LAGO Collaboration, a non-centralized and distributed collaborative network of more than 80 scientist from more than 25 institutions of 10 Latin American (currently Argentina, Bolivia, Brazil, Colombia, Ecuador, Mexico, Peru and Venezuela) and European (Spain) countries.

Also, detectors installed in various universities are used as a tool to teach students about particle and astroparticle physics, in particular by leading them to the measurement of the muon decay.

Today, LAGO is a network of ground-based Water Cherenkov Detector (WCDs) whose data are of interest for two different scientific communities:

- **Gamma Astronomy Community:** The LAGO WCDs installed at high altitude sites are sensitive to detect the effects of GRB. A significant number of LAGO detectors are over 3,000 m a.s.l. and three of them are above 4,300 m.
- **Space Weather:** The LAGO energy range covers a plethora of phenomena related to low-energy cosmic rays physics, from solar activity to space weather phenomena. Nowadays, the study of such phenomena is crucial to be studied because levels of radiation in the atmosphere and near-Earth space environment may be established.

Since the LAGO data analysis needs to take into account the influence of atmospheric effects, such as the pressure or the air temperature, on the flux of particles at the detector level, each LAGO WCD is equipped with several environmental sensors. These measurements represent an opportunity to provide environmental information to other communities, like ecologists studying the high altitude environments to correlate for possible climate change and global warming effects. Additionally, since the data in the LAGO repositories is open and freely accessible, it is used as motivation to train the general public -mainly the secondary school teachers and students- in statistical data analysis and related techniques, and raising the awareness of the general public in global warming and climate change impact in everyday life. This important citizen science initiative is one of the main objectives of the LAGO collaboration and is implemented in the so-called LAGO-CS (Citizen Science) program.

To properly analyze and exploit the measured data, computational capabilities must be available to the collaboration. This topic is being provisioned on a two-fold basis: customized codes that replicate the phenomena happened in the LAGO sites/detectors and data and metadata functionalities to be curated and managed.

With respect to the former, the CORSIKA code [1] used by LAGO covers three different energy ranges, which are different to the ones usually studied in other consortia such as the Pierre Auger Observatory[2].

On the other side, new challenges are emerging around the complex discovery environments in Data Centered Science, where trustworthiness and reproducibility of data are key requirements for new scientific findings. Experimental protocols are central part of any research and they are aimed to ensure the replicability of the results.

An increasingly frequent scenario is a researcher examining online the existing bibliography on a particular area. He/She finds several publications based on significant amount of data, registered, simulated or both and simultaneously is automatically redirected to the data used to produce them. Later on, the researcher access to the corresponding application employed to generate simulated data as well as the recorded raw/processed data. The new data, synthetic, measured or both (and the new paper if any) are stored on the Data Infrastructure and can be easily found making possible to start the cycle again.

This open access to data and/or applications will help to make reproducible the increasing complex workflow for producing scientific knowledge today.

2 Some hints about the LAGO related Physics

Nowadays, the LAGO Project is an extended Astroparticle Observatory at a continental scale, mainly oriented towards developing astroparticle physics at Latin America and doing basic research

in three areas: search for the high energy component of Gamma-ray bursts at high altitude sites, Space Weather phenomena, and Background Radiation at ground level [3]. The LAGO detection network consists in particle detectors deployed at ground level, spanning over different sites located at significantly different latitudes (currently planned from Mexico down the Antarctic region) and different altitudes (from sea level up to more than 5,000 meters over sea level), covering a large range of geomagnetic rigidity cut-offs and atmospheric absorption/reaction levels [4].

The current network of detectors is operated by the LAGO Collaboration, a non-centralized and distributed collaborative network of more than 80 scientist from institutions of nine Latin American countries (currently Argentina, Bolivia, Brazil, Colombia, Ecuador, Guatemala, Mexico, Peru and Venezuela) and Spain. Due to its proved reliability, high detection efficiency to all components present in atmospheric extensive showers, and low cost, water Cherenkov detectors (WCD) are currently used at LAGO sites [5].

The LAGO simulation chain consists of:

- Dynamic directional rigidity cut-off at each site $R(\text{Lat}; \text{Lon}; \theta; \phi; t)$
- Primary flux at the top of the atmosphere, i.e. CORSIKA simulations for each site $(\varphi; \lambda; h)$
 - Measured spectra for all nuclei $1 \leq Z_p \leq 26, 1 \leq A_p \leq 56$
 - $[R(\theta; \varphi) \times Z_p] \leq (E_p/\text{GeV}) \leq 10^6, 0^\circ \leq \theta \leq 90^\circ$
 - Integrated primary flux: $10^7 - 10^8 \text{ hour}^{-1} \text{ m}^{-2}$ (≥ 5 hours at each site)
- Secondary flux at detector level
- Detector response:
 - Fast and Simple LAGOFast detector simulation
 - Detailed GEANT4 model

3 LAGO simulations

The main code used in LAGO is CORSIKA. CORSIKA (COsmic Ray SIMulations for KAscade) is a program for detailed simulation of extensive air showers initiated by high energy cosmic ray particles. Protons, light nuclei up to iron, photons, and many other particles may be treated as primaries.

The particles are tracked through the atmosphere until they undergo reactions with the air nuclei or (in the case of unstable secondaries) decay. The hadronic interactions at high energies may be described by several reaction models alternatively: The VENUS, QGSJET, and DPMJET models are based on the Gribov-Regge theory, while SIBYLL is a minijet model. The neXus model extends far above a simple combination of QGSJET and VENUS routines. The most recent EPOS model is based on the neXus framework but with important improvements concerning hard interactions and nuclear and high-density effect. HDPM is inspired by findings of the Dual Parton Model and tries to reproduce relevant kinematical distributions being measured at colliders.

LAGO aims to study cosmic rays in the energy range 10GeV–100TeV. In this energy range there emerges phenomena related to the physics of low-energy cosmic rays, and also to solar activity and space weather environment. Nowadays it is crucial to study of these effects because it may establish levels of radiation in the atmosphere and near-Earth space environment.

To do so, the LAGO collaboration has implemented three specific and customized versions of CORSIKA that are to be executed on either local clusters or distributed environments such as Grid and Cloud.

In addition to CORSIKA, LAGO uses intensively other important codes: MAGNETOCOSMIC[6], GEANT4[7], ROOT[8] and specific self-designed statistical codes for data analysis, focusing most of the collaboration activities (research & outreach) on a data repository.

A fully dedicated Virtual Organization, called *lagoproject* is already integrated into the European Grid Infrastructure (EGI)¹ activities. The Grid implementation of CORSIKA was deployed in two 'flavors' being able to run by using GridWay Metascheduler² [9] or with a second approach

¹<http://www.egi.eu>

²<http://www.gridway.org/doku.php>

through a Catania Science Gateway interface[10]. For the former, massive calculations can be executed via the Montera [11], the GWpilot [12] or the GWcloud [13] frameworks

In the Science Gateway approach a user can seamlessly run a code on different infrastructures by accessing a unique entry point with an identity provision. He/she only has to upload the input data or use a PID to reference it and click on the run icon. The final result will be retrieved whenever the job will be ended. The underlying infrastructure is absolutely transparent to the user and the system decides on which sites and computing platform the code is performed.

4 The LAGOData e-infrastructure

Typically, each detector generates 150 GB of data per month and the entire collaboration generates 1.5 TB/month. The LAGO dataset not only refers to data measured by WCD detectors but also to data generated by simulation of cosmic rays phenomena in the aforementioned energy range. The CORSIKA particle flux simulations carried out generate 10GB/site and these synthetic data are also preserved in the data repository.

The low energy limit depends on the geomagnetic coordinates of the site, while the high energy limit is determined by the collection area at each site and is limited by statistics as the flux becomes lower and lower at higher energies (in general, the cosmic rays flux decreases by a factor of 1000 for an increase of 10 in the cosmic ray energy). This raw data collected by the LAGO detectors are shared through LAGOData[14], a platform conceived to promote data curation and sharing among LAGO collaborators, which is part of a more ambitious project, LAGOVirtual[15] a working environment which ensure access to the data recorded in all LAGO Sites and facilitate the analysis of such data.

4.1 Data curation through DSpace

Dspace is an open source software that enables sharing of many types of content, it is generally used for institutional repositories, providing basic functionality for saving, storing, and retrieving of digital content. DSpace was adopted for the LagoDATA repository, because it hosts Dublin Core metadata with a straightforward adaptability for non-native metadata schemes. It also supports two important interoperability protocols: OAI-PMH (Open Archive Initiatives Protocol for Metadata Harvesting³) and SWORD (Simple Webservice Offering Repository Deposit⁴). The OAI-PMH protocol at the LAGO repository allows the CHAIN-REDS Knowledge Base search engine to navigate into LAGO curated data.

It was important to overcome one of the most important DSpace limitations, i.e. its inability to upload/download multiple records. Dspace offers the possibility to upload the corresponding metadata through a command line option via an *import tool* by using *simple archive format* and including it in a separate way. A script to ingest data profiting from the above mentioned DSpace capability has been developed as well. With this scheme, it is worth mentioning that part of the metadata is the PID

4.2 The LAGOData metadata

The Dublin Core metadata element set is a standard for cross-domain information resource description, it is elaborated and sponsored by DCMI (Dublin Core Metadata Initiative⁵), the implementation of which makes use of XML. Dublin Core is Resource Description Framework based⁶ and comprises fifteen metadata elements: Title, Subject, Description, Source, Language, Relation, Coverage, Creator, Publisher, Contributor, Rights, Date, Type, Format, and Identifier. Despite this functionality is mostly centered on the Dublin Core metadata scheme, the additional non-native metadata can be configured as custom fields which are also stored, searched and displayed as the native ones.

The datasets are classified into three different types with their corresponding associated metadata: WCD, simulated, and calibration. Thus:

³<http://www.openarchives.org/pmh/>

⁴<http://swordapp.org/>

⁵<http://dublincore.org/>

⁶<http://www.w3.org/RDF/>

- WCD metadata scheme is: **data** corresponds to the version/type of the Digit/Analog electronic board; **site** contains the *name*, *latitude*, *longitude* and *height* of the detector; **voltage**, **level** and **sensor**;
- simulation metadata uses: **primary** described by the CORSIKA input file DATXXXX.dbase; **site** with the *latitude*, *longitude* and *height* of the ground point; **libraries** indicating which are the included CORSIKA libraries; **computation** describing the computational environment by unix command `uname -a`, `lsb_release -a`, `free` and `gcc -v`;
- calibration data refers to the calibration parameters used in the LAGO site.

4.3 PID and LAGOData

The main interface to register and manage PID services for European Research Communities is EPIC (European Persistent Identifiers Consortium⁷) which is based on the Handle System⁸ for the allocation and resolution of persistent identifiers. There are several compatible 'flavors' of PIDs. The most common is DOI PIDs⁹. DOI PIDs are more frequently used for publications while the EPIC PIDs cover a wider range of Digital objects.

The GRNET PID service¹⁰ enables the allocation, management and resolution of PIDs and has been employed to ensure the data persistence and reproducibility of the experiments. It supports the use of part identifiers as they are provided by the Handle system. Part identifiers can compute an unlimited number of handles on the fly, without requiring registering each separately.

4.4 SWORD and LAGOData

The SWORD (Simple Web-service Offering Repository Deposit) [16], based on the Atom Publishing Protocol (AtomPub), was first developed to standardize a deposit interface to digital repositories. Presently, it further extend the limited capabilities of AtomPub by supporting the whole deposit lifecycle, i.e. deposit, update, replace, and delete resources. Many interfaces of laboratory equipment allows automatic capture of results in an information system and SWORD permits to upload data directly into a repository, without human intervention, tagging as metadata how data was collected and the conditions in which were collected.

4.5 The DART challenge in LAGO

The Data Accessibility, Reproducibility and Trustworthiness (DART) initiative was launched by CHAIN-REDS¹¹ (Coordination and Harmonisation of Advanced e-infrastructure for Research and Education Data Sharing), an European Commission co-funded project focused on promoting and supporting technological and scientific collaboration across different communities in various continents. This initiative provided a set of interrelated tools and services, based on worldwide adopted standards, which made possible to easily/seamlessly access datasets, data/documents repositories and the applications that could generate and/or make use of them.

Trustworthiness can be associated to data curation, particularly on the quality of the metadata describing the experimental protocol and data provenance, while *reproducibility* and *replicability* is closely connected to the accessibility to data sources and the possibility to manipulate/analyze data contained in them.

CHAIN-REDS approach to data trustworthiness and reproducibility is based on the integration of computational resources and services, with three main cornerstones:

⁷<http://pidconsortium.eu>

⁸The Handle System (<http://www.ietf.org/rfc/rfc3650.txt>). This provides efficient, extensible, and secure resolution mechanism for unique and persistent identifiers of digital objects. The Handle System includes an open set of protocols (<http://hdl.handle.net/4263537/4086>), a namespace (<http://hdl.handle.net/4263537/4068>), and a reference implementation (<http://handle.net/download.html>) of the protocols.

⁹The Digital Object Identifier (DOI) System (<http://www.doi.org/>), a service operated by the International DOI Foundation (IDF), which provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers on digital networks.

¹⁰<http://epic.grnet.gr>

¹¹<http://www.chain-project.eu>

1. adoption of standards for data discoverability, provenance and recoverability: OAI-PMH¹² for metadata retrieval, Dublin Core¹³, as metadata schema, SPARQL¹⁴ for semantic web search and XML¹⁵ as potential standard for the interchange of data;
2. enablement of datasets authorships and user authentication with the corresponding assignments of specific roles on data services, which can be implemented by two strategies:
 - assignment of PIDs to name data in a unique and timeless manner, ensuring that future changes on URIs or internal organization of databases will be transparent to the user
 - implementation of federated identity provision, a secure, flexible and portable mechanism to access e-infrastructures worldwide, based on agreements and standards.
3. access to a plethora of computing power to analyze the retrieved data or to contrast them to simulations through an intuitive web-interface. Over the last years, Science Gateways have risen as an ideal tool to allow scientists across the world to seamlessly access different ICT-based infrastructures for research activities to support their day-by-day work and do better (and faster) research.

In other words: user identification; execution of distributed applications with a simple web interface; usage of Open Access Document Repositories and Data Repositories; and reproducibility of the experiments conform the DART challenge, where the whole research cycle is covered and can be seamlessly performed by non-expert users, hiding complex processes under simple interfaces and minimizing the need for learning new tools [18, 17].

5 Conclusion

The LAGO collaboration is working on studying cosmic rays in the energy range 10GeV–100TeV, which complements other astroparticle initiatives. In this energy range there emerges phenomena related to the physics of low-energy cosmic rays, and also to solar activity and space weather environment. Nowadays it is crucial to study these effects because they may establish levels of radiation in the atmosphere and near-Earth space environment. Thus the data repository (and the network of data repositories) will be of interest not only for LAGO and even the cosmic ray community, but useful for the solar physics and space climatology communities.

Such a research activity is being carried out by fostering collaboration in Latin America using experimental sites allocated along the continent as well as Cluster, Cloud and Grid Computing resources geographically distributed too.

Acknowledgment

The LAGO Collaboration is very thankful to the Pierre Auger Collaboration for its continuous support

References

- [1] D. Heck *et al.*, *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers*, Karlsruhe, Germany: Forschungszentrum Karlsruhe Report FZKA 6019, 1998.
- [2] The Pierre Auger Collaboration, *Correlation of the Highest-Energy Cosmic Rays with Nearby Extragalactic Objects*, *Science* **318** (5852), 938 (2007)
- [3] H. Asorey *et al.*, *The LAGO space weather program: Directional geomagnetic effects, background fluence calculations and multi-spectral data analysis*, in *The 34th International Cosmic Ray Conference*, vol. PoS(ICRC2015), 142 (2015)

¹²<http://www.openarchives.org/pmh/>

¹³<http://dublincore.org>

¹⁴<http://www.w3.org/2001/sw/wiki/SPARQL>

¹⁵<http://www.w3.org/XML/>

- [4] I. Sidelnik, *The sites of the Latin American giant observatory*, in The 34th International Cosmic Ray Conference, vol. PoS(ICRC2015), 665 (2015)
- [5] H. Asorey *et al.*, *LAGO: the Latin American Giant Observatory*, in The 34th International Cosmic Ray Conference, vol. PoS(ICRC2015), 247 (2015)
- [6] L. Desorgher, *MAGNETOCOSMICS: Geant4 application for simulating the propagation of cosmic rays through the Earth's magnetosphere*, available at <http://reat.space.qinetiq.com/septimes/magcos/>, 2004.
- [7] S. Agostinelli *et al.*, *Geant4 – A Simulation Toolkit*, Nuclear Instruments and Methods A **506**, 250 (2003)
- [8] R. Kumar and A. Tripathi, *Root: A data analysis and data mining tool from cern*, Casualty Actuarial Society E-Forum, 1 (2008).
- [9] E. Huedo *et al.*, *The gridway framework for adaptive scheduling and execution on grids*, Scalable Computing: Practice and Experience, **6** (3), 1, 2001.
- [10] R. Barbera, M. Fargetta and R. Rotondo, *A Simplified Access to Grid Resources by Science Gateways*, In The International Symposium on Grids and Clouds and the Open Grid Forum, Taipei, Taiwan, March 2011.
- [11] M. Rodríguez-Pascual *et al.*, *Montera: a framework for efficient execution of Monte Carlo codes on Grid infrastructures*, Computing and Informatics **32**, 113 (2013)
- [12] A.J. Rubio-Montero *et al.*, *hGWpilot: Enabling multi-level scheduling in distributed infrastructures with GridWay and pilot jobs*, Future Generation Computer Systems **45**, 25 (2015)
- [13] A.J. Rubio-Montero *et al.*, *User-Guided Provisioning in Federated Clouds for Distributed Calculations*, LNCS **9438**, 60 (2015)
- [14] L.A. Torres *et al.*, *Implementación de un repositorio de datos científicos usando dspace*, E-Colabora, **1** (2), 101 (2011)
- [15] R. Camacho *et al.*, *LAGOVirtual: A collaborative environment for the large aperture grb observatory*. In R. Mayo, H. Hoeger, L. Ciuffo, R. Barbera, I. Dutra, P. Gavillet, and B. Marechal, editors, *Proceedings of the Second EELA2 Conferencem Choroni Venezuela, Madrid España, 2009. EELA2, CIEMAT*.
- [16] S. Lewis *et al.*, *SWORD: Facilitating Deposit Scenarios*, D-Lib Magazine, **18** (1-2), (2012)
- [17] H. Asorey *et al.*, *Data Accessibility, Reproducibility and Trustworthiness with LAGO Data Repository*, in The 34th International Cosmic Ray Conference, vol. PoS(ICRC2015), 672 (2015)
- [18] M. Rodríguez-Pascual *et al.*, *A resilient methodology for accessing and exploiting data and scientific codes on distributed environments*, in *Procs. 2015 IEEE 18th International Conference on Computational Science and Engineering*, 319 (2015)